Balanced Truncation of a Stable Non-Minimal Deep-Submicron CMOS Interconnect

Amir Zjajo, Qin Tang, Michel Berkelaar, Nick van der Meijs

***Abstract*—As the widening of process variability in submicron CMOS technology calls for accurate timing models, their deployment requires well-controlled characterization techniques to cope with the complexity and scalability. In this context, model order reduction techniques have been used extensively to reduce the complexity of extracted interconnect circuits and to expedite fast and accurate circuit simulation. In the interconnect modeling, solving large-scale Lyapunov equations arises as a necessity in model order reduction techniques based on Balanced Truncation. In this paper, within this framework, dominant eigensubspaces of the product of the system Gramians are approximated directly. We construct orthogonal basis sets for the dominant subspaces of controllability and observability Gramians and perform eigenvalue decomposition to reduce the cost of singular value decomposition. As the experimental results indicate, the proposed approach can significantly reduce the complexity of interconnect, while retaining high accuracy in comparison to the original model.**

***Index Terms*—interconnect model, model order reduction, balanced truncation.**

I. Introduction

GATE and interconnect delay are critical issues in present day low power VLSI circuit design. As we are moving towards nanometer technology, variations in process, voltage, and temperature are increasing, causing significant uncertainty in the delay estimation [1] and greatly impacting the yield [2]. As a consequence, various statistical static timing analysis (SSTA) algorithms [3]-[5] have been proposed to compute the statistical variations of timing performance due to the underlying process parameters. Deriving an efficient characterization methodology and model order reduction (MOR) techniques that can provide parameterized interconnects and facilitate efficient logic stage delay calculation is one of the critical tasks. In an asymptotic waveform evaluation (AWE) algorithm [6] explicit moment matching was used to compute the dominant poles via Padé approximation. As the AWE method is numerically unstable for higher-order moment approximation, a more elegant solution to the numerical problem of AWE is to use projection-based MOR methods. In the Padé via Lanczos (PVL) method [7], the Lanczos process, which is a numerically stable method for computing eigenvalues of a matrix, was used to compute the Krylov subspace. In PRIMA [8] the Krylov subspace vectors are used to form the projector for the congruence transformation, which leads to passive models with the matched moments in the rational approximation paradigm. However, these methods are not efficient for circuits with many inputs and output terminals as the reducing cost are tied to the number of terminals; the number of poles of reduced models is also proportional to the number of terminals. Additionally, PRIMA-like methods do not preserve structure properties like reciprocity of a network.

Another approach to circuit-complexity reduction is to reduce the number of nodes in the circuits and approximate the newly added elements in the circuit matrix in reduced rational forms by approximate Gaussian elimination for RC circuits [9]. Alternatively, model order reduction can be performed by means of singular-value-decomposition (SVD) based approaches such as control-theoretical-based truncated balance realization (TBR) methods, where the weakly uncontrollable and unobservable state variables are truncated to achieve the reduced models [10]-[16]. The major advantage of SVD-based approaches over Krylov subspace methods lies in their ability to ensure the errors satisfying an a-priori upper bound [14]. Also, SVD-based methods typically lead to optimal or near optimal reduction results as the errors are controlled in a global way, although, for large scale problems, iterative methods have to be used to find an adequate balanced approximation (truncation). In this respect, ideas based on balanced reduction methods are significant since they offer the possibility to perform order selection during the computation of the projection spaces and not in advance. Typically in balanced reduction methods, there is a rapid decay in the Gramians eigenvalues. As a consequence these Gramians can be well approximated using low-rank approximations, which are used instead of the original. Accordingly, several SVD approaches approximate the dominant Cholesky factors (dominant eigensubspaces) of controllability and observability Gramians [11],[15]-[16] to compute the reduced model.

In this paper, we adjust the dominant subspaces projection model reduction (DSPMR) [11] and provide an approximate balancing transformation for circuits whose coefficient matrices are large and sparse such as in interconnect. The

This research was sponsored by the European Union and the Dutch government as part of the ENIAC/MODERN project.

The authors are with Circuits and Systems Group, Delft University of Technology, Mekelweg 4, 2628 CD, Delft, The Netherlands. (e-mail: amir.zjajo@ieee.org).

approach presented here produces orthogonal basis sets for the dominant singular subspace of the controllability and observability Gramians significantly reducing the complexity and computational costs of singular value decomposition, while preserving model order reduction accuracy and the quality of the approximations of the TBR procedure.

II. Adjusted Approximated Truncated Balance Realization Method

In the analysis of delay or noise in on-chip interconnect we study the propagation of signals in the wires that connect logic gates. These wires may have numerous features: bends, crossings, vias, etc., and are modeled by circuit extractors in terms of a large number of connected circuit elements: capacitors, resistors and more recently inductors. Given a state-space formulation of the interconnect model,

$$\begin{aligned} C(dx/dt) &= Gx(t) + Bu(t) \\ y(t) &= E^T x(t) \end{aligned} \tag{1}$$

where $C,G \in \mathcal{R}^{n\times n}$ are matrices describing the reactive and dissipative parts of the interconnect, respectively, $B \in \mathcal{R}^{n\times p}$ is a matrix that defines the input ports, $E \in \mathcal{R}^{p\times n}$ is matrix that defines the outputs, and $y(t) \in \mathcal{R}^q$ and $u(t) \in \mathcal{R}^p$, are the vectors of outputs and inputs, respectively, the model reduction algorithm seek to produce a similar system

$$\begin{aligned} \hat{C}d\hat{x}/dt &= \hat{G}\hat{x}(t) + \hat{B}u(t) \\ \hat{y}(t) &= \hat{E}^T \hat{x}(t) \end{aligned} \tag{2}$$

where $\hat{C},\hat{G} \in \mathcal{R}^{k\times k}$, $\hat{B} \in \mathcal{R}^{k\times m}$, $\hat{E} \in \mathcal{R}^{p\times k}$, of order k much smaller than the original order n, but for which the outputs $y(t)$ and $\hat{y}(t)$ are approximately equal for inputs $u(t)$ of interest. The Laplace transforms of the input output transfer functions

$$\begin{aligned} H(s) &= E^T (G + sC)^{-1} B \\ \hat{H}(s) &= \hat{E}^T (\hat{G} + s\hat{C})^{-1} \hat{B} \end{aligned} \tag{3}$$

are used as a metric for approximation accuracy if

$$\left\| H(s) - \hat{H}(s) \right\| < \varepsilon \tag{4}$$

for a given allowable error ε and an allowed domain of the complex frequency variable s, the reduced model is accepted as accurate.

Balanced truncation [10],[16], singular perturbation approximation [18], and frequency weighted balanced truncation [19] are model reduction methods for stable systems. Except for modal truncation each of the above methods is based either explicitly or implicitly on balanced realizations, the computation of which involves the solutions of Lyapunov equations

$$\begin{aligned} GXC^T + CXG^T &= -BB^T \\ G^T YC + C^T YG &= -E^T E \end{aligned} \tag{5}$$

where the solution matrices X and Y are controllability and observability Gramians.

The original implementation of balanced truncation [10] involves the explicit balancing of the realization (1). This procedure is dangerous from the numerical point of view because the balancing transformation matrix T tends to be highly ill-conditioned.

The square root method [16] is an attempt to cope with this problem by avoiding explicit balancing of the system. The method is based on the Cholesky factors of the Gramians instead of the Gramians themselves. In [20] the use of the Hammarling method was proposed to compute these factors. Recently, in [11] and [15] it has been observed that solutions to Lyapunov equations often have low numerical rank, which means that there is a rapid decay in the eigenvalues of the Gramians. Indeed, the idea of low-rank methods is to take advantage of this low-rank structure to obtain approximate solutions in a low-rank factored form. The principal outcome of these approaches is that the complexity and the storage are reduced from $O(N^3)$ flops and $O(N^2)$ words of memory to $O(N^2 r)$ flops and $O(Nr)$ words of memory, respectively, where r is the approximate rank of the Gramian ($r \ll N$). Moreover, approximating the Cholesky factors of the Gramians directly and using these approximations to provide a reduced model, has a comparable cost to that of the popular moment matching methods. It requires only matrix-vector products and linear solvers.

For large systems with a structured transition matrix, this method is an attractive alternative because the Hammarling method can generally not benefit from such structures. In the original implementation this step is the computation of exact Cholesky factors, which may have full rank. We formally replace these (exact) factors by (approximating) low rank Cholesky factors [11],[15]. The iterative procedure approximates the low rank Cholesky factors Z_X and Z_Y with $r_X, r_Y \ll n$, such that $Z_X Z_X^H \approx X$ and $Z_Y Z_Y^H \approx Y$, where H is Hermitian (complex-conjugate) matrix. Note that the number of iteration steps i_{max} needs not be fixed a priori. However, if the Lyapunov equation should be solved as accurate as possible, correct results are usually achieved for low values of stopping criteria that are slightly larger than the machine precision. Let

$$Z_Y^H Z_X = U_Y \Sigma U_X^H \tag{6}$$

be SVD of $Z_Y^H Z_X$ of dimension $N \times m$. The cost of this decomposition including the construction of U is $14Nm^2 + O(m^3)$ [21].

To avoid this, in this paper we perform eigenvalue decomposition

$$(Z_Y^H Z_X)^H Z_Y^H Z_X = U_Y \Lambda U_X^H \tag{7}$$

Comparing (7) with (6) shows that the same matrix U_X is constructed and that

$$(Z_Y^H Z_X U_X)^H Z_Y^H Z_X U_Y = \Lambda = \Sigma^H \Sigma \tag{8}$$

This algorithm requires Nm^2 operations to construct $(Z_Y^H Z_X)^H Z_Y^H Z_X$ and $Nmn + O(m^3)$ operations to obtain $Z_Y^H Z_X U_X \Sigma^{-1}$ for $n \times n$ Σ.

The balancing transformation matrix T is used to define the matrices $S_X=T_{(1:k)}$ and $S_Y=T^T_{(1:k)}$. If $\sigma_k \neq \sigma_{k+1}$, the reduced order realization is minimal, stable, and balanced, and its Gramians are equal to $diag(\sigma_1,\ldots,\sigma_k)$. The balancing transformation matrix can be obtained as

$$S_X = Z_X U_X \Sigma^{-1/2} \quad S_Y = Z_Y U_Y \Sigma^{-1/2} \tag{9}$$

then, under a similarity transformation of the state-space model, both parts can be treated simultaneously after a transformation of the system (C,G,B,E) with a nonsingular matrix $T \in \mathcal{R}^{n\times n}$ into a balanced system

$$\widehat{C} = S_X C S_Y^H \quad \widehat{G} = S_X G S_Y^H \quad \widehat{B} = S_Y^H B \quad \widehat{E} = E S_X \tag{10}$$

In this algorithm we assume that $k \leq r$ ($rank\ Z_Y^H Z_X$). Note that SVDs are arranged so that the diagonal matrix containing the singular values has the same dimensions as the factorized matrix and the singular values appear in non-increasing order.

III. Experimental Results

The proposed method and all sparse techniques have been implemented in Matlab. All the experimental results are carried out on a PC with an Intel Core 2 Duo CPU running at 2.66 GHz and with 3 GB of memory. To characterize the timing behavior, a lookup table-based library is employed which represents the gate delay and output transition time as a function of input arrival time, output capacitive load, and several independent random source of variation for each electrical parameter (i.e., R and C). In each case, both driver and interconnect are included for the stage delay characterizations. The analytical delay distribution obtained using the quadratic interconnect model in 45 nm CMOS technology is illustrated in Figure 1. The nominal value of the total resistance of the load and the total capacitance is chosen from the set 0.15kΩ-1kΩ and 0.4pF-1.4pF, respectively.

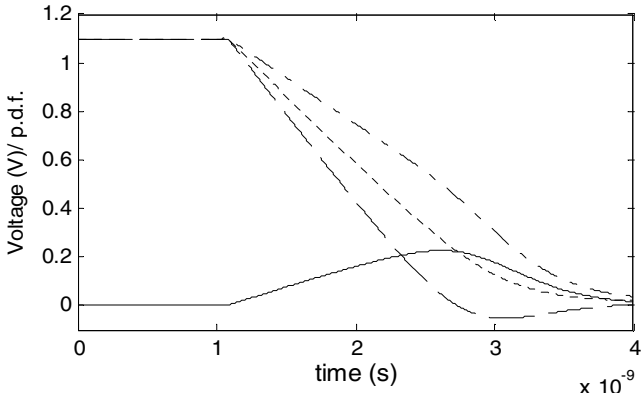

Figure 1: Analytical delay distribution in 45 nm CMOS technology. Solid line illustrates delay variance.

The sensitivity of each given data to the sources of variation is chosen randomly, while the total σ variation for each data is chosen in the range of 10% to 30% of their nominal value. The scaled distribution of the sources of variation is considered to have a skewness of 0.5, 0.75, and 1.

For model order reduction we consider a *RC*-chain with 2002 capacitors and 2003 resistors. In Figure 2 and Figure 3 the convergence history with respect to the number of iteration steps for solving the Lyapunov equation is plotted. For the tolerances at a residual norm of about the same order of magnitude, convergence is obtained after 40 and 45 iterations, respectively. The *cpu*-time needed to solve the Lyapunov equations according to the related tolerance for solving the shifted systems inside the iteration is 2.7 seconds.

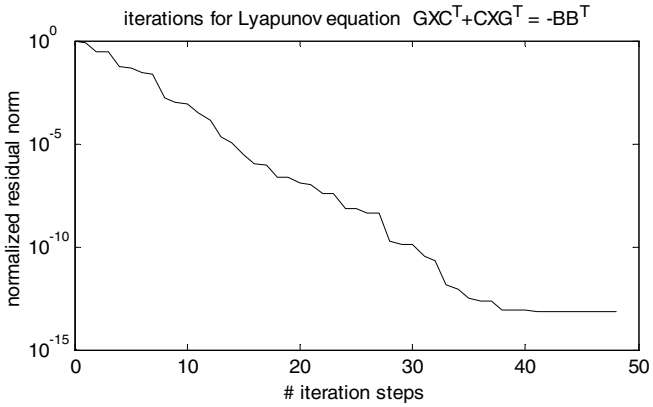

Figure 2: Convergence history of residual forms. The convergence is obtained after 40 iterations.

Note further that saving iteration steps means that we save large amounts of memory-especially in the case of multiple input and multiple output systems where the factors are growing by p columns in every iteration step. When very accurate Gramians (e.g. low rank approximations to the solutions) are selected, the approximation error of reduced system as illustrated in Figure 4 is very small compared to the Bode magnitude function of the original system. The lower two curves correspond to the highly accurate reduced system; the proposed model order reduction technique delivers a system of lower order, and the upper two denote k=20 reduced orders. The frequency response plot is obtained by computing the singular values of the transfer function $H(j\omega)$, which is the frequency response (4) evaluated on the imaginary axis (Figure 5). The error plot is the frequency response plot of the singular values of the error system as a function of ω.

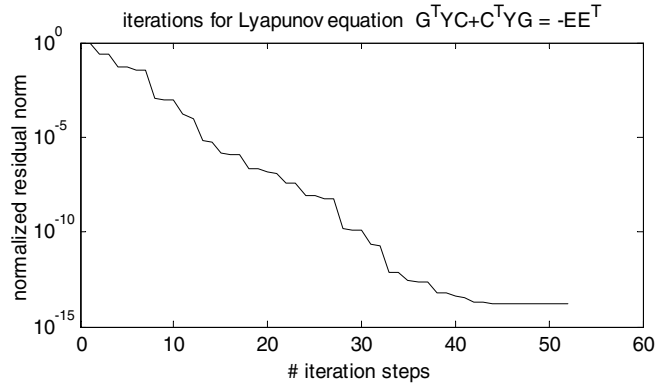

Figure 3: Convergence history of residual forms. The convergence is obtained after 45 iterations.

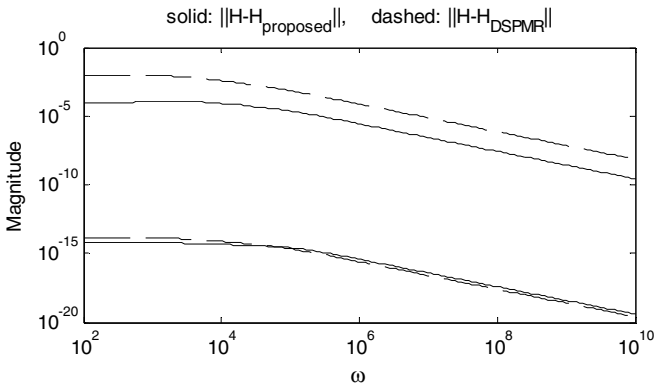

Figure 4: The Bode magnitude plot of the approximation errors.

The reduced order is chosen in dependence of the descending ordered singular values $\sigma_1,\sigma_2,\ldots\ \sigma_r$, where r is the rank of factors which approximate the system Gramians. For n variation sources and l reduced parameter sets, the full parameter model requires $O(n^2)$ simulation samples and thus has a $O(n^6)$ fitting cost. On the other hand, the proposed parameter reduction technique has a main computational cost attributable to the $O(n+l^2)$ simulations for sample data collection and $O(l^6)$ fitting cost significantly reducing the required sample size and the fitting cost.

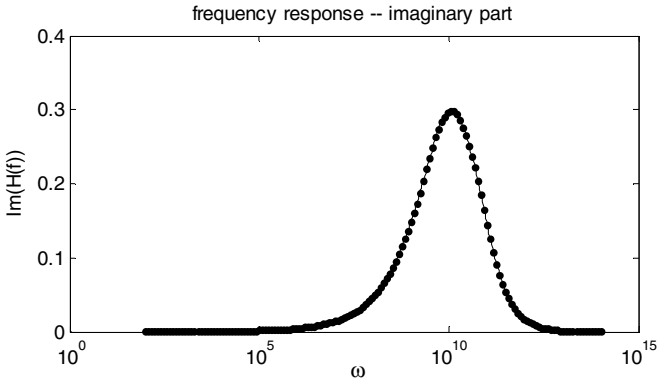

Figure 5: Frequency response of the interconnect model.

IV. CONCLUSION

This paper presents an efficient methodology for interconnect model reduction based on adjusted dominant subspaces projection. By adopting the parameter dimension reduction techniques, interconnect model extraction can be performed in the reduced parameter space, thus provide significant reductions on the required simulation samples for constructing accurate models. Extensive experiments are conducted on a large set of random test cases, showing very accurate results.

REFERENCES

[1] C. Forzan, D. Pandini, "Statistical static timing analysis: A survey," *Integration, The VLSI Journal,* vol. 42, no. 3, pp. 409-435, 2009

[2] S.R. Nassif, "Modeling and analysis of manufacturing variations," *Proceedings of IEEE Custom Integrated Circuit Conference*, pp. 223-228, 2001

[3] L. Zhang, W. Chen, Y. Hu, A. Gubner, C. Chen, "Correlation-preserved non-Gaussian statistical timing analysis with quadratic timing model," *Proceedings of IEEE Design Automation Conference*, pp. 83-88, 2005

[4] V. Veetil, D. Sylvester, D. Blaauw, "Efficient Monte Carlo based incremental statistical timing analysis," *Proceedings of IEEE Design Automation Conference*, pp. 676-681, 2008

[5] Q. Tang, A. Zjajo, M. Berkelaar, N. van der Meijs, "RDE-based transistor-level gate simulation for statistical static timing analysis," *Proceedings of IEEE Design Automation Conference*, pp. 787-792, 2010

[6] L. T. Pillage, R. A. Rohrer, "Asymptotic waveform evaluation for timing analysis," *IEEE Transaction on Computer-Aided Design of Integrated Circuits and Systems*, vol. 4, pp. 352-366, 1990

[7] P. Feldmann, R.W. Freund, "Efficient linear circuit analysis by Pade approximation via the Lanczos process," *IEEE Transaction on Computer-Aided Design of Integrated Circuits and Systems*, vol. 14, pp. 639-649, 1995

[8] A. Odabasioglu, M. Celik, L. Pileggi, "PRIMA: Passive reduced-order interconnect macromodeling algorithm," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, pp. 645-654, 1998

[9] P. Elias, N. van der Meijs, "Including higher-order moments of RC interconnections in layout-to-circuit extraction," *Proceedings of IEEE Design, Automation and Test in Europe Conference*, pp. 362-366, 1996

[10] B. C. Moore, "Principal component analysis in linear systems: controllability, observability, and model reduction," *IEEE Transaction on Automatic Control*, vol. 26, pp. 17-31, 1981

[11] J. Li, J. White, "Efficient model reduction of interconnect via approximate system Grammians," *Proceedings of IEEE International Conference on Computer Aided Design*, pp. 380-384, 1999

[12] J. R. Phillips, L. Daniel, L. M. Silveira, "Guaranteed passive balancing transformations for model order reduction," *Proceedings of IEEE Design Automation Conference*, pp. 52-57, 2002

[13] J. R. Phillips, L. M. Silveira, "Poor man's TBR: a simple model reduction scheme," *Proceedings of IEEE Design, Automation and Test in Europe Conference*, pp. 938-943, 2004

[14] W.F. Arnold, A.J. Laub, "Generalized eigenproblem algorithms and software for algebraic Riccati equation," *Proceedings of IEEE*, vol. 72, pp. 1764-1754, 1984

[15] T. Penzl, "A cyclic low-rank Smith method for large sparse Lyapunov equations," *SIAM Journal on Scientific Computing*, vol. 21, pp. 1401-1418, 2000

[16] M.G. Safonov, R.Y. Chiang, "A Schur method for balanced-truncation model reduction," *IEEE Transactions on Automatic Control*, vol. 34, pp. 729-733, 1989

[17] A.J. Laub, M.T. Heath, C.C. Paige, R.C. Ward, "Computation of system balancing transformations and other applications of simultaneous diagonalization algorithms," *IEEE Transactions on Automatic Control*, vol. 32, pp. 115-122, 1987

[18] K.V. Fernando, H. Nicholson, "Singular perturbational model reduction of balanced systems," *IEEE Transactions on Automatic Control*, vol. 27, pp. 466-468, 1982

[19] D. Enns, "Model reduction with balanced realizations: an error bound and a frequency weighted generalization," *Proceedings of IEEE Conference on Decision and Control*, pp. 127-132, 1984

[20] M.S. Tombs, I. Postlethwaite, "Truncated balanced realization of stable, non-minimal state-space systems," *International Journal of Control*, vol. 46, pp. 1319-1330, 1987

[21] G. Golub, C. van Loan, *Matrix computations*, Johns Hopkins University Press, Baltimore MD, 1996